Avatars signeurs : quels défis pour la production de contenu en Langues des Signes ?

Sylvie Gibet

Avatars signeurs : quels défis pour la production de contenu en Langues des Signes ?

Sylvie Gibet
IRISA
Université Bretagne Sud
Vannes, France
sylvie.gibet@univ-ubs.fr

***Résumé*—Cet article explore les défis relevés par les systèmes permettant de générer des contenus en langues des signes au moyen d'avatars signeurs 3D. Après avoir passé en revue les avatars signeurs existants et leurs spécificités, à la fois au niveau de la représentation linguistique et du système d'animation qu'ils proposent, nous décrivons quelques mécanismes de formation des signes et des énoncés, à la lumière de procédés de spatialisation et d'iconicité caractérisant ces langues visuo-gestuelles. Nous présentons ensuite les défis pour la génération de signes à partir de texte.**

***Mots clés*—avatar signeur, langues des signes, génératon, texte vers signes, spatialisation, iconicité**

I. Introduction

Les langues des signes (LS) sont des langues à part entière qui caractérisent l'identité et la culture sourde. Elles appartiennent à la famille des langues dites *visuo-gestuelles* pour lesquelles l'information est émise par les gestes et perçue par le système visuel. Ainsi, les personnes sourdes développent avec la pratique de cette langue une dextérité dans leur gestuelle et dans leur perception visuelle, une acuité de représentation de l'espace et une expressivité qui s'exprime dans leurs mouvements manuels, corporels et leurs expressions faciales. C'est pourquoi on qualifie ces langues de "multi-canal", l'information étant véhiculée sur les différents canaux représentés par les mains, le corps, le visage et la direction du regard. Ainsi, tous les segments corporels participent à diffuser le message. Les mains constituent bien sûr le principal vecteur d'émission de l'information, mais il faut leur adjoindre certains gestes et mouvements corporels ainsi que les expressions faciales qui sont primordiales pour qualifier affectivement une entité, une action ou une phrase, voire pour exprimer la négation ou l'interrogation.

Comme toute langue, les LS possèdent une capacité d'expression et d'abstraction qui s'appuie sur une structure linguistique propre avec son vocabulaire, sa grammaire et sa sémantique. De plus, elles intègrent dans les processus de formation et de flexion des signes des mécanismes d'iconicité et de spatialisation que l'on ne retrouve dans aucune langue orale. Enfin, les messages signés exploitent les caractéristiques propres aux mouvements, à la fois au niveau de leur génération et de leur perception. Ainsi, un énoncé signé à un rythme similaire à celui de langues parlées est traduit en un flux continu de mouvements qui sont perçus comme étant naturels et compréhensibles.

La plupart des outils numériques à destination des personnes sourdes signantes, que ce soit le téléphone mobile avec vidéo intégrée, les dispositifs de visioconférence, ou Internet et les réseaux sociaux, sont particulièrement adaptés pour les personnes pratiquant les LS puisque les applications utilisées sont essentiellement basées sur une communication visuelle. Cependant, si la vidéo est le média le plus partagé par les sourds, elle ne permet pas de garantir l'anonymat et impose des contraintes fortes au niveau du stockage et du transport d'information. La production automatique de messages en LS et la visualisation au moyen d'avatars signeurs, définis comme des personnages virtuels en 3D capables de s'exprimer en langues des signes, semblent constituer une réponse alternative appropriée, en permettant à la fois la réduction des informations stockées, l'anonymisation ainsi que la programmabilité des personnages virtuels pour éditer et produire de nouveaux énoncés. Parmi les applications exploitant des avatars signeurs on peut citer plus particulièrement celles qui visent à améliorer : (i) l'éducation en LS ; (ii) l'apprentissage des LS ; (iii) l'accessibilité à la connaissance pour les personnes pratiquant les LS (web, traducteur texte-vers-signe) ; (iv) la conception de vidéo-book, en particulier pour les enfants.

Dans cet article, nous donnons un aperçu des principales technologies existantes pour la génération de contenu en LS, et présentons certains défis qui restent à relever pour traduire un texte en signes tout en respectant les mécanismes linguistiques propres aux LS et en utilisant des avatars signeurs animés. Quelques exemples en Langue des Signes Française (LSF) illustrent le propos.

II. Avatars signeurs existants

Nous commençons par passer en revue certaines des technologies utilisées pour animer les avatars signeurs. La figure II présente par ordre chronologique quelques avatars signeurs existants.

Les premiers travaux sur la phonologie du signe ont donné lieu à différents types de représentations. Parmi celles-ci, les travaux de Stokoe [6] ont abouti à la description de l'ASL (American Sign Language) sous la forme d'une combinaison d'unités élémentaires constituant les signes : l'emplacement du signe, la forme de la main et son mouvement. L'une des hypothèses de base repose sur le fait que deux signes distincts peuvent être différenciés lorsque l'un seulement

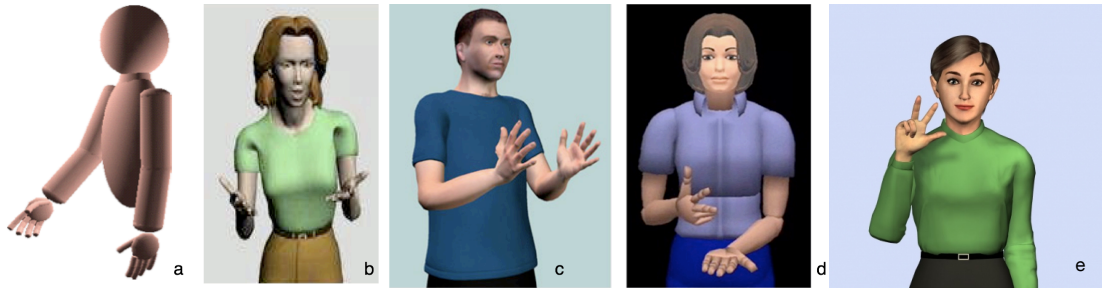

FIGURE 1. Quelques avatars signeurs classés par ordre chronologique : (a) le système GESSYCA [1] (b) Elsi [2] (c) Guido du projet européen eSign [3] (d) le signeur virtuel de l'université de New-York City [4] (e) Paula [5]

des paramètres constitutifs est modifié (paires minimales). Un dictionnaire de l'ASL a été constitué à partir de cette représentation. Poursuivant les travaux de Stokoe, d'autres paramètres qui participent à la formation et la distinction des signes ont été identifiés. Ils incluent l'orientation de la main et les paramètres non manuels (en particulier l'expression faciale) [7]. Ces éléments phonologiques sont agencés et synchronisés spatialement et temporellement pour former des signes et des énoncés en LS.

Plus tard, le système *HamNoSys* (*Hamburg Notation System* [8]) a proposé un système de notation qui reprend les paramètres précédents et transcrit les signes de manière linéaire en utilisant les symboles informatiques Unicode. Avec une approche linguistique de l'ASL, Liddell & Johnson ont défini un système phonétique [9] qui s'appuie sur le modèle Posture-Détention-Transition-Shift (*PDTS*) distinguant sur chaque canal – configuration de la main HC, orientation FA, placement PL, caractéristiques non manuelles NM – des éléments statiques et des éléments dynamiques transitionnels.

Plus récemment, en linguistique computationnelle, les gestes des LS ont été décrits au moyen de formalismes allant de scripts à des langages gestuels dédiés. Le langage *SiGML* [10] basé sur *HamNoSys* a été développé pour générer les animations d'avatars 3D. Ce langage a ensuite été étendu en incorporant le modèle *PDTS* de Johnson & Liddell. Le langage impératif *QualGest* [1] s'appuie sur une logique phonologique et une description spatiale qualitative pour définir une grammaire de formation des signes. Le langage de modélisation *Azee* est quant à lui basé sur une formalisation géométrique [11].

Ces langages de scripts permettent de décrire des signes ou des énoncés de manière très analytique et précise. Cependant, la spécification de nouveaux signes peut être très fastidieuse. On peut noter que la plupart des langages de spécification intègrent au sein de leur formalisme des éléments temporels explicites, c'est le cas notamment de *SIGML*, de *EMBRscript* [12] ou de *Azee* dans lesquels les postures clés de l'avatar sont spécifiées à des instants pré-déterminés. Par contre, le langage *QualGest* se base sur une formalisation du temps implicite, la synchronisation entre les mouvements des différents articulateurs étant gérée au niveau des moteurs d'animation.

Le passage de la spécification des signes à la génération de mouvement a donné lieu à des travaux qui visent à traduire une description textuelle en une séquence de commandes gestuelles directement interprétables par des moteurs d'animation. La plupart des travaux existants concernent des méthodes de synthèse "pure", qui consistent à calculer par interpolation la suite des postures de l'avatar à partir de la spécification de postures clés. C'est le cas du système d'animation d'avatar en ASL *Paula* (see Fig. II) de DePaul University basé sur un moteur d'animation multi-pistes. Il exploite la représentation phonétique *PDTS* de Jonhson & Liddell comme un patron pour spécifier manuellement les postures clés [5] et créer ainsi des signes. D'autres systèmes d'animation d'avatar s'appuient sur des techniques de cinématique inverse pour générer des mouvements. C'est le cas du système (*JASigning*) qui intègre le moteur d'animation *AnimGen* permettant la création de signes spécifiés à partir du langage (*SiGML*) [14], ou du système *GesSyCA* associé au langage *QualGest*.

Si ces systèmes d'animation, couplant un langage de script à un moteur d'animation utilisant de la synthèse pure, permettent d'atteindre des objectifs de précision et de contrôle fin des mouvements, à la fois corporels, manuels et faciaux, ils donnent généralement lieu à des mouvements robotisés. De plus, ils permettent de créer un nombre de signes limité. En effet, construire un vocabulaire de signes et d'énoncés en LS par de telles méthodes peut s'avérer être une tâche chronophage en temps de spécification. Enfin, la gestion du temps reste complexe à mettre en oeuvre, à la fois au niveau des signes (gestion de la synchronisation entre composants des signes) et des transitions entre signes (gestion de la coarticulation).

Une alternative à ces systèmes de synthèse consiste à développer des méthodes d'animation basées données. Dans ce cas, les mouvements d'une personne signante sont capturés par des techniques de capture de mouvement qui permettent d'enregistrer simultanément les mouvements des mains, du corps et des expressions faciales (voir Figure 2). Ainsi, les systèmes *SignCom* [15] et *Sign3D* [13] ont permis d'animer des avatars en LSF avec des mouvements naturels et réalistes à partir de mouvements réels de signeurs. Dans le cadre de ces systèmes, deux bases de données ont été constituées : (i) une base de données de mouvements bruts dans laquelle les

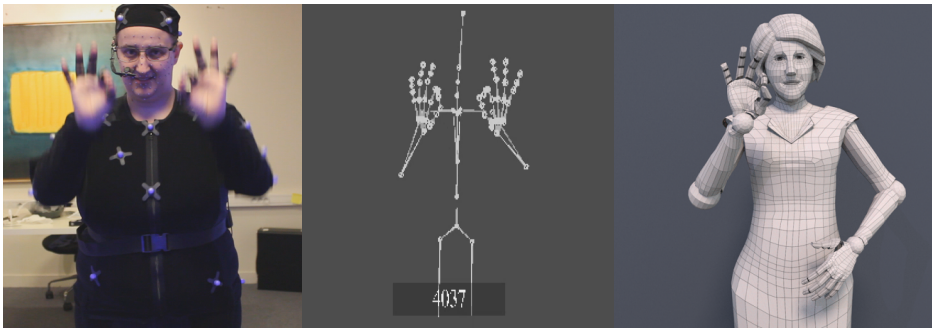

FIGURE 2. Chaîne de capture de mouvement dans le projet Sign3D [13]

mouvements capturés sont stockés sous la forme de postures squelettiques caractérisées par des transformations appliquées aux articulations, et (ii) une base de données sémantique qui met en correspondance les annotations multi-niveaux des signes et les mouvements. Le système d'animation est conçu à partir d'un principe de synthèse concaténative multi-pistes, chaque piste étant associée à un ensemble de contrôleurs dédiés (animation faciale, direction du regard, animation corporelle ou manuelle). Ce système a permis d'éditer et de construire de nouveaux énoncés en LSF : (i) par remplacement de signes ou de groupes de signes ; (ii) par instantiation de schémas syntaxiques stéréotypés ; (iii) ou par remplacement d'éléments phonologiques (configurations manuelles, mouvements des mains, du torse, de la tête, expressions faciales) [16].

Les méthodes basées données permettent de produire des animations d'avatars en LS qui sont très fluides et crédibles. Elles permettent de rejouer des séquences relativement longues (discours en LS), mais également d'éditer les phrases pré-enregistrées afin de produire de nouveaux énoncés. Cependant la manipulation et l'adaptation des mouvements au contexte requiert la prise en compte de processus linguistiques élaborés afin de garder la cohérence des contenus produits, à la fois au niveau des animations et de la compréhensibilité des LS.

III. CARACTÉRISTIQUES DES LANGUES DES SIGNES

Les langues des signes remettent en question les frontières habituelles des théories linguistiques associées aux langues orales. Ceci est principalement dû au fait qu'elles utilisent l'information gestuelle et visuelle, contrairement aux langues orales qui utilisent le canal audio-oral. Cette spécificité est à l'origine de l'omniprésence des mécanismes iconiques et spatiaux dans les langues des signes. L'iconicité telle que définie par Cuxac [17] met en jeu des processus par lesquels le locuteur va rendre iconique l'expérience vécue, imaginée ou exécutée. Elle est caractérisée par le lien de ressemblance plus ou moins étroit entre les entités du monde réel, le référent et le signe qui s'y rapporte. Cuxac propose ainsi une théorie "de grande iconicité", dans laquelle deux sortes de signes coexistent lors d'activités discursives : (i) les signes qui "disent en montrant", et les signes "standards" (dans leur forme de citation), sans visée illustrative. Les mécanismes engendrés par l'iconicité structurent et modulent les langues des signes. Une autre spécificité des LS concerne la difficulté de séparer les différents niveaux linguistiques – phonétiques, phonologiques, morphologiques, grammaticaux et sémantiques – qui sont propres aux langues orales [18]. Ainsi, dans les LS, toute modification effectuée au niveau des composants constitutifs des signes (composants phonétiques ou phonologiques, par analogie aux langues orales) peut potentiellement altérer le sens d'un signe ou celui de la phrase elle-même.

Nous évoquons ci-après de manière non exhaustive quelques défis relatifs à la production des signes et des énoncés signés.

IV. PRODUCTION DES SIGNES ET DES ÉNONCÉS EN LANGUES DES SIGNES

Les langues des signes exigent précision et rapidité dans leur exécution, mais en même temps, une imperfection dans la réalisation des signes ou une mauvaise synchronisation peuvent modifier le contenu sémantique de la phrase. Inversement, une modification, même subtile, de l'énoncé se traduit souvent par une modification sensible des mouvements produits, à la fois d'un point de vue de la sélection et de l'organisation des signes, que des procédés flexionnels appliqués aux signes. Ces procédés diffèrent, suivant qu'il s'agit de signes à visée non illustrative (signes standards) ou des signes à visée illustrative. C'est pourquoi nous séparons ces deux catégories de signes, avant de nous intéresser à la construction d'énoncés par composition des différents mécanismes.

A. Formation des signes à visée non illustrative

La formation des signes à visée non iconique (ou signes standards) nécessite la combinaison spatio-temporelle des composants phonologiques décrits ci-dessus. Ces signes sont généralement toujours exécutés de la même manière à un emplacement spécifique de l'espace entourant le signeur, encore appelé espace de signation, et la modulation de ces signes provient uniquement de la variabilité au niveau des

FIGURE 3. Les signes AIMER / NE-PAS-AIMER en LSF à gauche, et DONNER / PRENDRE à droite : les trajectoires des mains sont inversées

mouvements exécutés. Si ces signes ne sont pas soumis aux processus de flexion, leur réalisation requiert toutefois une grand précision, la modification d'un composant phonologique engendrant un sens différent, comme par exemple le signe NATUREL qui devient le signe PAS-BESOIN en modifiant la configuration manuelle, l'emplacement et les mouvements étant inchangés. D'un point de vue temporel, il est nécessaire également de respecter des règles de synchronisation entre les éléments composant le signe.

B. Formation des signes et énoncés à visée illustrative

Les mécanismes de spatialisation et d'iconicité sont très présents dans les langues des signes. Nous explorons ci-après quelques procédés identifiés en LSF. La flexion opérée sur ces signes ou séquences de signes, qui se caractérise par la modification d'une ou plusieurs composantes phonologiques, conduit à modifier la qualité ou le sens du signe ou de l'énoncé.

Spatialisation

Plusieurs mécanismes relatifs à la spatialisation sont listés ci-dessous :

- Positionnement. Pour exprimer les positions absolues ou relatives entre entités signifiées, les langues des signes ont recours aux positions dans la scène de signation. Le signeur utilise ainsi l'espace en positionnant les entités, animées ou non, présentes dans sa narration. Il est à noter que ce positionnement peut être absolu ou relatif.
- Pointage. La désignation d'un emplacement se fait souvent par pointage de l'index (ou autre configuration manuelle).
- D'autres techniques permettent de préciser une localisation : l'orientation du regard, de la tête ou du buste pendant l'exécution du signe, ou le décalage du signe dans l'espace.
- Les verbes directionnels sont ceux qui s'accordent avec l'agent et le patient. Ils incorporent dans leur réalisation les pronoms personnels et sont tels que les positions et orientations initiale et finale dépendent de l'agent et du patient par rapport auxquels le signe est fléchi. Par exemple le verbe DONNER peut se décliner suivant différentes lignes directionnelles et sens du mouvement en fonction de l'agent et du patient visé ("Je te donne" ou "Tu me donnes"). Pour le verbe DEMANDER, il n'y a pas de mouvement de la main, mais les configurations et orientations de la main changent selon les pronoms personnels.

Description de formes

La description des formes est au centre de l'iconicité présente dans les LS. Ci-dessous quelques procédés propres à la LSF :

- Flexion des verbes directionnels. Outre la flexion suivant les agents et patients, certains verbes directionnels transitifs peuvent être fléchis en fonction du complément d'objet direct. Dans ce cas, la forme de la main est modifiée. Par exemple, "Je te donne une lettre" ou "Je te donne un livre" sont réalisés en LSF de la même façon, hormis la configuration manuelle qui change pour représenter soit la lettre, soit le livre.
- Le transfert de taille ou de forme permet de représenter la taille ou la forme d'entités ou de personnes. Il concerne, soit des descriptions spatiales statiques (par exemple PETIT-BATEAU ou GROS-BATEAU), soit des descriptions dynamiques impliquant un suivi de trajectoire (PETIT-BOL, GRAND-BOL).
- Les proformes statiques représentent des entités animées (personnes, véhicules) et sont caractérisées par un nombre restreint de configurations. Elles évitent de nommer plusieurs fois une entité et rendent plus efficace le référencement de ces entités dans l'espace. Par exemple, la proforme PERSONNE peut être rapidement positionnée dans la scène de narration. De plus, la personne peut être représentée dans différentes positions (debout, assise ou allongée), ce qui conduit à des configurations manuelles différentes. De même on peut représenter facilement plusieurs personnes dans un espace (autour d'une table par exemple) ou dans une salle de conférence.

Aspects propres à la dynamique des mouvements

- Trajectoires. Des mécanismes spatiaux et temporels précis concernent les trajectoires des mains au sein des signes, qui ne sont pas seulement des transitions, mais prennent la forme d'une ligne, d'un arc, ou d'une forme plus complexe telle qu'une ellipse, une spirale, etc. Par exemple, le signe AIMER est représenté par un

FIGURE 4. L'avion décolle.

mouvement d'arc vers le haut. Il est possible d'inverser le mouvement de ce signe pour produire le sens NE-PAS-AIMER (Figure 3, gauche). De la même manière, inverser le signe DONNER peut produire le signe PRENDRE (Figure 3, droite).

— Proformes dynamiques. Certains comportements ou démarches peuvent également être représentés par des proformes dynamiques. C'est le cas par exemple lorsque l'on veut décrire la démarche d'un animal (un oiseau, un ours ou un lion par exemple). La forme de la main est modifiée pour représenter la forme de la patte de l'animal ainsi que sa démarche plus ou moins lourde. Il est possible de plus d'indiquer par le mouvement des mains la qualité du mouvement (souplesse, légèreté).

— Dynamique des mouvements. La dynamique temporelle et physique des mouvements peut également modifier la signification des signes. Ainsi, les signes CHAISE et S'ASSEOIR possèdent les mêmes configurations manuelles, les mêmes trajectoires spatiales mais ont des dynamiques différentes. Il est à noter également que la façon dont les contacts sont exécutés (de manière effleurée ou frappée) modifie le sens des signes.

Expressions faciales

Les expressions faciales sont primordiales en LS. Elles ne sont pas seulement une information liée à la qualité de ce qui est exprimé (émotion, prosodie), mais elles constituent des informations objectives qui participent à la sémantique de la phrase.

— Certaines expressions faciales expriment des adverbes ("Le vent souffle fort") ou des adjectifs ("Un homme mince ou gros").

— D'autre expressions faciales expriment les affects qui concernent tout ou partie de la phrase. Une mauvaise interprétation de ces émotions montrées délibérément peut altérer le sens de la phrase. Par exemple si l'on relate un accident en affichant un sourire, cela peut être mal interprété.

— Les expressions faciales donnent aussi des informations sur l'aspect clausal de la phrase. Une négation peut s'exprimer par un signe à part entière ou bien par une expression faciale (sourcils froncés indiquant la négation en fin de phrase). De même la phrase interrogative se distingue de la phrase déclarative par l'expression interrogative du visage. Dans certaines phrases conditionnelles il est possible également d'utiliser des expressions faciales spécifiques ("S'il pleut, je reste à la maison").

— Prise de rôle. Enfin la prise de rôle caractérise le fait d'incarner le rôle d'un personnage. Dans ce cas, le visage exprime certaines caractéristiques physiques ou psychologiques du personnage dont il endosse le rôle.

Mouvements corporels

Les mouvements corporels (par opposition aux mouvements manuels) sont aussi porteurs d'information.

— Ainsi, en LSF, le tronc légèrement penché en avant peut indiquer une action réalisée dans le futur.

— L'orientation du buste peut-être utilisée pour le changement de rôle (changement de référencement de personnages).

— Enfin, dans la description de formes, il est possible d'utiliser les différents plans ou niveaux de l'objet décrit en se penchant en avant.

C. Composition des différents éléments dans le discours

L'ensemble des mécanismes précédemment décrits sont utilisés dans des phrases discursives (histoires, dialogues, etc.) en langues des signes. Les signes standards et les signes à visée illustrative sont agencés dans les énoncés, soit séquentiellement soit en modifiant des informations sur un ou plusieurs canaux phonologiques. Dans l'exemple donné à la Figure 5, la phrase "J'aime les jus de fruits" est transformée en la phrase "Je n'aime pas le jus d'orange". Le mouvement du buste ainsi que celui du bas du corps et du bras gauche sont conservés. Par contre, les mouvements de la tête et du bras droit, ainsi que l'expression faciale sont modifiés.

Il est à noter qu'en fonction des signes, les mouvements des mains peuvent être réalisés avec une main seule (la main dominante) ou bien avec les deux mains. Les mouvements des

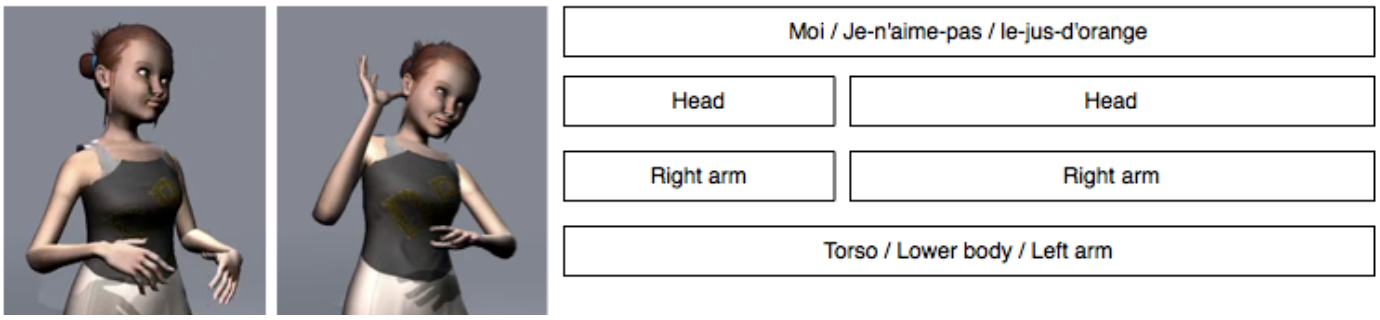

FIGURE 5. Combinaison de trois signes : moi / je-n'aime-pas / le-jus-d'orange.

mains peuvent être totalement symétriques par rapport à l'un des trois plans – sagittal, longitudinal, horizontal –, ou bien alternés par rapport à l'un de ces plans, ou encore ils peuvent être réalisés de manière dissymétrique avec une main dominée qui établit la base du signe et une main dominante qui effectue l'action principale, comme dans l'exemple "L'avion décolle" de la Figure 4.

Dans un récit tiré du corpus de Cuxac, mettant en scène deux chiens de morphologies très différentes (un boxer et un chien loup), les deux protagonistes du récit sont décrits à l'aide de transferts de forme sans que le signe standard [CHIEN] ne soit utilisé [17]. Les attributs des deux chiens protagonistes du récit sont caractérisés par le museau, les oreilles, les bajoues du boxer ainsi que ses pattes. Ils sont utilisés pour raconter une histoire dans un processus narratif exploitant les mécanismes spatiaux et iconiques des LS [19].

V. DÉFIS POUR LES AVATARS SIGNEURS

La Figure 6 recense les étapes nécessaires pour animer un avatar signeur à partir d'une représentation textuelle dans une langue donnée (génération *Texte-vers-Signe*, haut). Elle permet de localiser les défis encore rencontrés aujourd'hui pour les avatars signeurs. Si l'on se réfère aux outils existants présentés à la section II, l'étape 2 constitue le niveau de représentation linguistique computationnelle d'une LS (appelée ici *Pivot-LS*) et les étapes 3-4 le moteur d'animation avec le passage de la spécification symbolique vers la production d'un flux continu de postures de l'avatar, à partir d'un ensemble de contrôleurs de mouvement. La génération *Texte* vers *Pivot-LS* concerne la traduction d'un texte écrit dans une langue orale vers un texte dans un langage "pivot" représentant une langue des signe spécifique. Nous décrivons ci-après les défis relatifs à ces différents étapes.

A. Langages de représentation des LS : Pivot-LS

Afin de tenir compte des mécanismes d'inflexion linguistiques, il est nécessaire à ce niveau d'incorporer dans la représentation du langage une paramétrisation des signes qui tienne compte de la variation spatiale et temporelle des éléments les constituant. Par exemple, la phrase "Je te donne un livre", peut s'exprimer par : DONNER("je", "tu", LIVRE), expression qui peut être interprétée par un mouvement de la main correspondant au signe DONNER, depuis la localisation du pronom "je" à celle du pronom "tu", et une configuration manuelle qui est celle du signe LIVRE. La difficulté à ce niveau consiste à identifier un nombre limité d'expressions du langage et de paramètres associés qui couvre l'ensemble des mécanismes de flexion des signes. Un autre défi consiste à intégrer dans le langage des contraintes spatiales et temporelles qui pourront ensuite être interprétées par les contrôleurs d'animation.

B. Animation de l'avatar 3D

Contrôleurs mixtes Parmi les technologies permettant d'animer des avatars signeurs à partir d'une spécification textuelle, deux grandes approches ont été abordées (voir section II) : la première consiste à construire de toute pièce des avatars robotisés, pour lesquels on maîtrise toute la chaîne de production dans les moindres détails, depuis la spécification analytique des éléments linguistiques de base, leur agencement au moyen d'une logique procédurale ou à base de règles, jusqu'à la production d'une séquence de postures clés de l'avatar 3D. La seconde approche part du matériau de base qu'est le mouvement, et extrait les caractéristiques linguistiques par un processus d'annotation (manuelle ou automatique), pour permettre ensuite de construire des animations par composition des éléments annotés. Les deux types de contrôle (basé données et synthèse pure) peuvent être combinés. Il est possible en effet de substituer aux processus de synthèse "pure" des processus d'extraction et de recherche de postures ou de mouvements dans une base de données préalablement étiquetée, ou bien de remplacer des méthodes de synthèse par des méthodes de génération basées données s'appuyant sur un apprentissage automatique.

Les expressions de l'étape 2 peuvent alors être interprétées comme des commandes des contrôleurs du mouvement associés à un groupe d'articulateurs spécifique. Par exemple, pour la phrase "Je te donne", on obtient l'activation de deux contrôleurs : (i) le premier est un contrôleur de type IK (cinématique inverse) qui s'applique à la chaîne articulée du bras droit et qui prend en paramètres deux positions (celles des deux pronoms "je" et "tu") ; (ii) le second est un contrôleur de cinématique directe s'appliquant à la main droite et qui prend

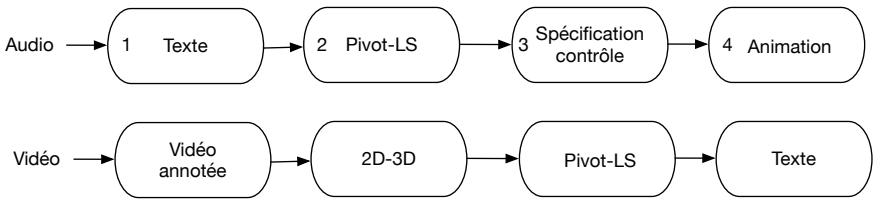

FIGURE 6. Chaînes de traitement ; en haut : Texte-vers-Signe ; en bas : Signe-vers-Texte

en paramètres deux configurations manuelles HC_1 et HC_2 (en début et en fin de mouvement).

Coordination multi-articulateurs L'un des principaux défis concerne la coordination multi-articulateurs et la synchronisation des mouvements générés sur chacune des pistes (soit par extraction dans la base de données, soit par synthèse) de façon à respecter les *patterns* spatio-temporels des signes. En effet, pour que les signes générés paraissent plausibles, il est nécessaire qu'ils vérifient des lois du mouvement humain [20]. D'autre part les mouvements relatifs des groupes d'articulateurs doivent respecter des schémas de synchronisation réalistes : par exemple la configuration manuelle doit être atteinte avant que la main ait atteint son objectif cible. D'un point de vue temporel, il peut être nécessaire de compresser ou de dilater des portions de mouvements de façon à ce que les règles de synchronisation soient vérifiées. Cette synchronisation se répercute également aux mouvements secondaires apparaissant dans certains signes.

Coarticulation Par ailleurs, il est primordial de gérer la coarticulation, et ceci à plusieurs niveaux, sur chacune des pistes et entre les signes de façon à tenir compte du contexte passé et futur de chaque signe dans la séquence générée.

Adaptation morphologique La plupart des avatars signeurs utilisent les données signées d'un seul signeur. L'adaptation morphologique à d'autres avatars signeurs passe par des processus d'adaptation morphologique (*retargeting*) permettant par exemple de transférer les animations vers d'autres personnages (homme, femme, enfant, voire animal). Alors que l'interaction avec le sol ou les objets dans l'environnement entraîne des contraintes dures qui conduisent à des problèmes et des procédures d'optimisation difficiles, les contraintes des LS peuvent être plus diffuses ou exprimées de manière qualitative (par exemple "le pouce doit toucher la paume de la main"). Les algorithmes traitant de ces contraintes de haut niveau pourraient être extrêmement intéressants, tant sur le plan numérique (relâchement des degrés de liberté tout en optimisant) que d'un point de vue de l'utilisabilité. Enfin, une phase de planification peut également être nécessaire pour éviter les auto-collisions. La combinaison de la cinématique inverse et d'algorithmes de planification pourrait être utilisée [21], ainsi que des approches hybrides [22]. Cependant, les algorithmes en temps réel pour cette catégorie de problèmes restent à définir. Il est à noter la place particulière de l'animation des mains en LS qui nécessite une grande précision et l'évitement des interpénétrations, d'où la prise en compte de contraintes spatiales dans les processus d'optimisation [23] (Figure 7).

FIGURE 7. Animation des mains : prise en compte de contraintes spatiales

C. Corpus de données

Les données sont au coeur des technologies et méthodes employées pour les avatars signeurs. Trois catégories de données sont disponibles : la vidéo, la capture de mouvement et les annotations. Plusieurs questions se posent pour la définition du corpus. La première concerne le compromis entre étendue et profondeur du corpus. Si l'objectif est de disposer d'un lexique qui couvre un large domaine, comprenant plusieurs thématiques, un corpus étendu sera privilégié. Si, au contraire, l'objectif est d'avoir un vocabulaire limité et de le réutiliser dans différentes phrases, alors on choisira l'approche en profondeur. Dans ce cas, de nombreuses instances des mêmes signes avec variations seront considérées dans le vocabulaire prédéfini. La deuxième question concerne la nature des variations elles-mêmes qui doivent être incluses dans le corpus pour l'édition et la synthèse. Plusieurs niveaux de variation des signes peuvent être considérés, incluant : (i) des variations du contexte, par exemple en faisant varier les prédécesseurs et successeurs d'un même signe, facilitant ainsi l'étude de la coarticulation ; (ii) des modulations de signes au niveau de leurs composants élémentaires, en modifiant par exemple les emplacements, les configurations manuelles ou les mouvements des mains ; (iii) des variations spatiales permettant l'étude spécifique de certains mécanismes flexionnels tels que les transferts de forme ou de taille, les pointages, les déclinaisons des verbes directionnels selon les pronoms et les compléments d'objet ; (iv) des variations du style ou de la prosodie qui induisent des modifications cinématiques des mouvements produits (plus ou moins rapides, fluides ou saccadés, etc.). Enfin, une préoccupation essentielle relative à la construction du corpus est la qualité actée ou spontanée des mouvements produits par les acteurs signants.

D. Génération Texte-vers-Signe et Signe-vers-Texte

Les systèmes actuels de traduction automatique d'une langue orale vers une autre laissent entrevoir la possibilité de traduire automatiquement une langue orale vers une langue des signes (transformation $1 \rightarrow 2$). Les méthodes d'apprentissage profond (*deep learning*) devraient faciliter cette étape. Cependant, l'absence de système d'écriture communément accepté pour les LS ne permet pas de disposer de suffisamment d'information mettant en correspondance un texte dans une langue orale et sa transcription en LS. Avec le peu de corpus parallèles disponibles, seuls des systèmes de traduction à base de règles sont actuellement envisagés.

Les données vidéo étant plus faciles à acquérir que les données de mouvements capturés, il est possible de disposer à court terme de gros volumes de données vidéo. Cela permettrait de développer la chaîne de traitement Signe-vers-Texte (Figure 6, bas) en exploitant des méthodes de *deep learning*. Transformer des données vidéo 2D en des données 3D reste toutefois un enjeu important pour les LS qui demandent une grande précision des mouvements manuels et des expressions faciales. Il est peut-être possible de s'affranchir du passage vers le squelette 3D et de produire directement à partir de vidéos les informations textuelles. Notons que la reconstruction du squelette 3D permettrait de constituer des bases de données conséquentes associant vidéo et MoCap, celles-ci pouvant être exploitées pour la reconnaissance de signes à partir de vidéos ou pour la synthèse de mouvements à partir de texte.

Il subsiste également la question de l'alignement vidéo / texte qui n'est pas résolue, préférablement dans le langage pivot ou à défaut dans la langue orale.

VI. Conclusion

Dans cet article, nous avons mis en évidence un ensemble non exhaustif de mécanismes linguistiques propres aux langues des signes et décrit certains défis technologiques pour la production de signes et d'énoncés au moyen de langages dédiés et d'avatars signeurs animés. Si certains travaux permettent d'ores et déjà d'appréhender ces mécanismes, à la fois d'un point de vue modélisation linguistique et développement informatique de modèles pour les implémenter, la grande variabilité propre à ces langues et la complexité des mécanismes de flexion, reposant sur la mise en oeuvre de processus spatio-temporels dédiés, ouvrent des voies de recherche encore peu explorées.

Dans un futur proche, la possibilité de capturer de grands volumes de données et l'avènement des méthodes d'apprentissage automatique profond vont permettre de concevoir des systèmes autonomes de synthèse *Texte-vers-Signe* ou *Signe-vers-Texte* en s'appuyant directement sur des données vidéos annotées. Plus largement, cela ouvre des perspectives vers des systèmes de traduction automatique des langues orales vers les langues des signes, ou vice-versa.

Références

[1] S. Gibet, T. Lebourque, and P. Marteau, "High level specification and animation of communicative gestures," *Journal of Visual Languages and Computing*, vol. 12, pp. 657–687, 2001.

[2] M. Filhol, A. Braffort, and L. Bolot, "Signing avatar : Say hello to elsi !„" in *Proc. of Gesture Workshop 2007*, ser. LNCS, Lisbon, Portugal, Jun. 2007.

[3] J. R. Kennaway, J. R. W. Glauert, and I. Zwitserlood, "Providing signed content on the internet by synthesized animation," *ACM Trans. Comput.-Hum. Interact.*, vol. 14, no. 3, p. 15, 2007.

[4] M. Huenerfauth, L. Zhao, E. Gu, and J. Allbeck, "Evaluation of american sign language generation by native asl signers," *ACM Trans. Access. Comput.*, vol. 1, no. 1, pp. 1–27, 2008.

[5] J. McDonald, R. Wolfe, J. Schnepp, J. Hochgesang, D. G. Jamrozik, M. Stumbo, L. Berke, M. Bialek, and F. Thomas, "An automated technique for real-time production of lifelike animations of american sign language," *Universal Access in the Information Society*, vol. 15, no. 4, pp. 551–566, 2016.

[6] W. C. Stokoe, *Semiotics and Human Sign Language*. Walter de Gruyter Inc., 1972.

[7] R. Battison, *Lexical borrowing in American sign language.* ERIC, 1978.

[8] S. Prillwitz and H. Z. für Deutsche Gebärdensprache und Kommunikation Gehörloser, *HamNoSys : Version 2.0 ; Hamburg Notation System for Sign Languages ; An Introductory Guide*. Signum-Verlag, 1989.

[9] R. E. Johnson and S. K. Liddell, "A segmental framework for representing signs phonetically," *Sign Language Studies*, vol. 11, no. 3, pp. 408–463, 2011.

[10] R. Elliott, J. R. Glauert, J. Kennaway, I. Marshall, and E. Safar, "Linguistic modelling and language-processing technologies for avatar-based sign language presentation," *Universal Access in the Information Society*, vol. 6, no. 4, pp. 375–391, 2008.

[11] M. Filhol, J. McDonald, and R. Wolfe, "Synthesizing sign language by connecting linguistically structured descriptions to a multi-track animation system," in *International Conference on Universal Access in Human-Computer Interaction*. Springer, 2017, pp. 27–40.

[12] A. Heloir and M. Kipp, "Real-time animation of interactive agents : Specification and realization," *Applied Artificial Intelligence*, vol. 24, no. 6, pp. 510–529, 2010.

[13] S. Gibet, F. Lefebvre-Albaret, L. Hamon, R. Brun, and A. Turki, "Interactive editing in french sign language dedicated to virtual signers : requirements and challenges," *Universal Access in the Information Society*, vol. 15, no. 4, pp. 525–539, 2016.

[14] R. Kennaway, J. R. Glauert, and I. Zwitserlood, "Providing signed content on the internet by synthesized animation," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 14, no. 3, p. 15, 2007.

[15] S. Gibet, N. Courty, K. Duarte, and T. Le Naour, "The signcom system for data-driven animation of interactive virtual signers : Methodology and evaluation," in *ACM Transactions on Interactive Intelligent Systems*, vol. 1, no. 1, 2011.

[16] S. Gibet, "Building french sign language motion capture corpora for signing avatars," in *Workshop on the Representation and Processing of Sign Languages : Involving the Language Community, LREC 2018*, Miyazaki, Japan, May 2018.

[17] C. Cuxac, *La langue des signes française (LSF) : les voies de l'iconocité (French) [French Sign Language : the iconicity ways]*, ser. Faits de langues. Ophrys, 2000.

[18] A. Millet and A. Morgenstern, *Grammaire descriptive de la langue des signes française : dynamiques iconiques et linguistique générale*. UGA Editions, 2019.

[19] B. Lenseigne and P. Dalle, "Using signing space as a representation for sign language processing," in *Gesture in Human-Computer Interaction and Simulation*. Berlin, Heidelberg : Springer, 2006, pp. 25–36.

[20] S. Gibet, J. Kamp, and F. Poirier, "Gesture analysis : Invariant laws in movement," in *Gesture-Based Communication in Human-Computer Interaction*, ser. LNCS, Springer, vol. 2915, 2003, pp. 1–9.

[21] D. Bertram, J. Kuffner, R. Dillmann, and T. Asfour, "An integrated approach to inverse kinematics and path planning for redundant manipulators," in *Int. Conf. on Robotic and Automation*, 2006, pp. 1874–1879.

[22] L. Zhang, M. C. Lin, D. Manocha, and J. Pan, "A hybrid approach for simulating human motion in constrained environments," *Computer Animation and Virtual Worlds*, vol. 21, no. 3–4, pp. 137–149, 2010.

[23] T. L. Naour, N. Courty, and S. Gibet, "Skeletal mesh animation driven by few positional constraints," *Computer Animation and Virtual Worlds*, vol. 30, no. 3-4, 2019.